

Original article

1  
2  
3  
4 **INSIGHT: A Tool for Fit-for-Purpose Evaluation and Quality**  
5 **Assessment of Observational Data Sources for Real World**  
6 **Evidence on Medicine and Vaccine Safety**  
7  
8  
9  
10

11 Vjola Hoxhaj<sup>1</sup> ([v.hoxhaj@umcutrecht.nl](mailto:v.hoxhaj@umcutrecht.nl))  
12 Constanza L. Andaur Navarro<sup>1</sup> ([c.l.andaurnavarro@umcutrecht.nl](mailto:c.l.andaurnavarro@umcutrecht.nl))  
13 Judit Riera-Arnau<sup>1,2</sup> ([j.rieraarnau@umcutrecht.nl](mailto:j.rieraarnau@umcutrecht.nl))  
14 Roel JHJ Elbers<sup>1</sup> ([roel@hotmail.com](mailto:roel@hotmail.com))  
15 Ema Alsina<sup>1</sup> ([palsinaaer@gmail.com](mailto:palsinaaer@gmail.com))  
16 Caitlin Dodd<sup>1</sup> ([caitlindodd@gmail.com](mailto:caitlindodd@gmail.com))  
17 Miriam CJM Sturkenboom<sup>1</sup> ([m.c.j.sturkenboom@umcutrecht.nl](mailto:m.c.j.sturkenboom@umcutrecht.nl))  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28

29 <sup>1</sup> Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht University,  
30 Utrecht, The Netherlands.

31 <sup>2</sup> Clinical Pharmacology Service, Vall d'Hebron Hospital Universitari, Vall d'Hebron Barcelona Hospital Campus,  
32 Universitat Autònoma de Barcelona, Barcelona, Spain.

33  
34  
35 **Word count:** 3705/4000  
36

37 **Keywords:** data quality assessment, quality checks, level checks, real-world data, common data  
38 model, pregnancy, drug safety, multi-database studies.  
39

40 **Corresponding author:** Constanza L. Andaur Navarro, [c.l.andaurnavarro@umcutrecht.nl](mailto:c.l.andaurnavarro@umcutrecht.nl)  
41 Julius Centre for Health Sciences and Primary Care, Universiteitsweg 100, P.O. Box 85500, 3508 GA,  
42 Utrecht, The Netherlands

**NOTE:** This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

43 **ABSTRACT**

44 **Objective.** To describe the development of INSIGHT, a real-world data quality tool to assess  
45 completeness, consistency, and fitness-for-purpose of observational health data sources.

46 **Material and Methods.** We designed a three-level pipeline with data quality assessments  
47 (DQAs) to be performed in ConcePTION Common Data Model (CDM) instances. The pipeline  
48 has been coded using R.

49 **Results.** INSIGHT is an open-source tool that identifies potential data quality issues in CDM-  
50 standardized instances through the systematic execution and summary of over 588  
51 configurable DQAs. Level 1 focuses on compliance with the ConcePTION CDM specifications.  
52 Level 2 evaluates the temporal plausibility of events and uniqueness of records. Level 3  
53 provides an overview of distributions, outliers, and trends over time. The DQAs are run locally  
54 and assessed centrally by a data quality revisor together with the data access provider's  
55 representatives.

56 **Discussion.** INSIGHT aligns with recent conceptual frameworks that identify five dimensions  
57 of data quality: reliability, extensiveness, coherence, timeliness, and relevance. Data quality  
58 is the sum of several internal and external features of the data and while DQAs provide  
59 reassurance about fitness-for-purpose for secondary-use data sources, improvements in data  
60 collection and generation stages are essential to reduce bias, misclassification, and  
61 measurement errors, thereby enhance overall data quality for Real World Evidence.

62 **Conclusion.** INSIGHT aims to support clinical and regulatory decision-making for medicines  
63 and vaccines by evaluating the quality of observational health data sources to support fit for  
64 purpose assessment. Assessing and improving data quality will enhance the reliability and  
65 quality of the generated evidence.

66

67 Word count: 246/250

## 68 INTRODUCTION

69 Regulatory agencies and healthcare professionals acknowledge and underline the  
70 significance of Real-World Data (RWD) in informing clinical decisions and shaping public  
71 health policies, especially for populations underrepresented in clinical trials, such as pregnant  
72 women.<sup>1,2</sup> Pharmacoepidemiologists have used RWD since the 90s, and worked in distributed  
73 networks to expedite the generation of Real-World evidence (RWE) for regulatory decision  
74 making. In such networks, multiple data access providers (DAPs) collaborate by handling  
75 heterogeneous data and using the state-of-the-art approach to generate RWE. This involves  
76 the use of a common protocol, a common data model and common analytics with tools, which  
77 have been improving over the past 15 years.<sup>3</sup> Comprehensive analyses of these approaches  
78 have been published.<sup>4</sup>

79 The lack of knowledge regarding the use and safety of medicines and vaccines during  
80 pregnancy and lactation is widely recognized. Studies indicate that 70-90% of women are  
81 exposed to a prescription medicine during pregnancy.<sup>5</sup> However, the process of determining  
82 the teratogenic status of a novel medication currently takes 27 years.<sup>6</sup> To close this evidence  
83 gap, the IMI ConcePTION project was established ([www.imi-conception.eu](http://www.imi-conception.eu)), bringing together  
84 88 stakeholders across Europe with the aim to create a learning healthcare ecosystem to  
85 support future studies on safety of medicines and vaccines during pregnancy and lactation.  
86 ConcePTION uses RWD to generate Real World Evidence (RWE) for childbearing-aged,  
87 pregnant, and lactating women, and their offspring.

88 Before conducting any analysis using RWD, it is imperative to ensure the quality of the data  
89 and Data Quality Assessments (DQAs) should be conducted to evaluate whether data sources  
90 are fit for purpose prior to study initiation. ConcePTION along with other research networks  
91 such as US Food and Drug Administration (FDA) Sentinel, PCORnet, and the Observational  
92 Health Data Science and Informatics program (OHDSI) perform DQAs on the data loaded into  
93 common data models (CDM) used by these networks for distributed analyses.<sup>7-9</sup> The  
94 ConcePTION CDM allows for syntactic harmonization of heterogeneous data sources, and  
95 enables fitness-for-use and fitness-for-purpose evaluations.<sup>10</sup> As part of the ConcePTION  
96 project, we have developed INSIGHT, a generic RWD tool to perform DQAs on ConcePTION  
97 CDM-standardized data sources ensuring data quality and the generation of reliable evidence  
98 for pharmacoepidemiologic studies. This tool aligns with Kahn's data quality framework which  
99 encompasses three main quality dimensions: *conformance*, *completeness*, and *plausibility*.<sup>11</sup>

100 In this article, we describe INSIGHT, an open-source tool that comprises three levels of DQAs  
101 and we outline step-by-step its design and execution. Furthermore, we present an overview  
102 of the DQA workflow, including processes, visualizations, and responsible parties. We show

103 examples of detected inconsistencies and errors, which serve as proof-of-concept,  
104 highlighting the tool’s effectiveness in identifying and addressing data quality issues.

105 **METHODS**

106 As part of a common analytics and distributed analytics framework, Data Access Providers  
107 (DAPs) are required to convert their local data sources into the ConcePTION CDM using an  
108 Extract-Transform-Load (ETL) process. The ConcePTION CDM encompasses demographics,  
109 observation periods, medicines and vaccines exposure, events, procedures, mother-child  
110 linkage, visits, and clinical observations. The current version 2.2 includes 16 tables ([Box 1](#)).  
111 Additional details can be found elsewhere.<sup>10</sup>

<b>Box 1. ConcePTION CDM tables</b>		112
<b>Routine healthcare data</b>	VISIT_OCCURRENCE	113
	EVENTS	114
	MEDICINES	115
	PROCEDURES	116
	VACCINES	117
	MEDICAL_OBSERVATIONS	118
<b>Surveillance</b>	EUROCAT	119
	SURVEY_ID	120
	SURVEY_OBSERVATIONS	121
	EUROCAT	122
<b>Curated tables</b>	PERSONS	123
	OBSERVATION_PERIODS	124
	PERSON_RELATIONSHIPS	125
	PRODUCTS	126
	CDM_SOURCE	127
<b>Metadata</b>	METADATA	128
	CDM_SOURCE	129
	INSTANCE	130
	PRODUCTS	131

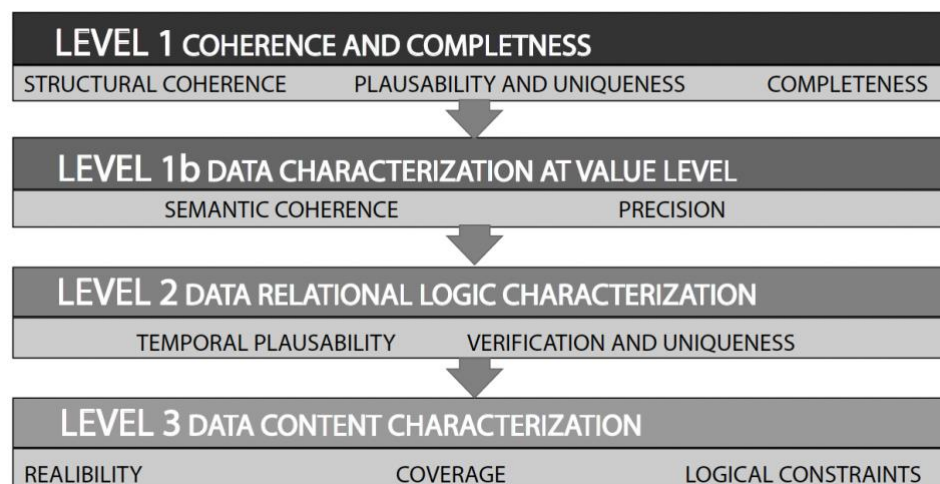
133 **ConcePTION Common data model**

134 When working on studies with the ConcePTION CDM, DAPs are asked to load their data into  
135 a standardized data structure using the source values. This ensures that the ConcePTION  
136 CDM is syntactically (ensuring consistency in the structure) harmonized. The ConcePTION  
137 CDM pipeline does not require semantic harmonization (translation/mapping of source value

138 to common vocabularies), which can often be time consuming. Instead, mapping is conducted  
139 as part of the study script, making it a transparent and flexible process. Once the data sources  
140 are syntactically standardized, DQAs are performed to allow assessment of data quality and  
141 fitness for purpose.

## 142 **Objectives of Data Quality Assessments**

143 Various conceptual data quality frameworks have been published over the last years.<sup>12</sup> For  
144 the development of INSIGHT, we focused on the harmonized terminology for data quality  
145 dimensions put forth by Khan et al., and the quality indicators established by the US FDA  
146 Sentinel initiative, OHDSI, and the EUROCAT indicators for population-based healthcare data  
147 sources.<sup>11,13</sup> By incorporating these frameworks, we aimed to create a comprehensive and  
148 robust data quality verification pipeline for projects using the ConcePTION CDM. In [Figure 1](#),  
149 we show the hierarchy of DQAs in INSIGHT followed by a brief description.



**Figure 1.** Hierarchy and dimensions of data quality assessment in INSIGHT

150 Level 1 data quality indicators provide insight on the completeness and compliance of the ETL  
151 process with the ConcePTION CDM specifications, ensuring format, structural and relational  
152 coherence. In addition, the plausibility of the data and the uniqueness of records is addressed,  
153 maintaining data integrity as well as identifying duplication errors.

154 Level 1b data quality indicators provide a list of all extracted and loaded non-date variables  
155 across the ConcePTION CDM tables. The primary objective is to provide DAPs, principal  
156 investigators, programmers, and analysts a comprehensive overview of the values of each  
157 non-date variable to allow for study-specific variable definitions. This level aims to enhance  
158 the understanding of the data content, facilitating semantic harmonization in the study scripts.

159 Level 2 data quality indicators provide an overview of the logical relationship and integrity of  
160 values within and between variables and tables. Its purpose is to verify the temporal plausibility  
161 by examining records that occur outside the recorded person-time, identifying observations  
162 related to a person ID that is not present in the PERSONS table, among other similar  
163 inconsistencies.

164 Level 3 data quality indicators provide distributions of population, diagnoses, medicines,  
165 vaccines, lifestyle factors, pregnancy, and temporal trends over calendar time for each specific  
166 variable. The primary objective is to allow for inspection of temporal changes in population,  
167 follow-up, medicines, vaccines, and disease rates. These indicators can be compared  
168 between instances and between DAPs, but also against external benchmarks to verify their  
169 fitness for purpose.

170 To ensure that data quality indicators can be inspected, it is important to present the results  
171 in a format that facilitates their understanding and sharing. To achieve this, HTML reports with  
172 the presentation of the data quality indicators were developed. These reports contain summary  
173 tables which allow for a concise representation of data quality indicators and graphs that  
174 provide a visual representation of trends and patterns. INSIGHT has been implemented using  
175 R scripts, which automate the running process of DQAs. The INSIGHT DQA is an iterative  
176 process, each level can be rerun until the required quality is attained or all constraints are  
177 noted.

## 178 **Design of Data Quality Indicators**

### 179 *Level 1: Coherence and Completeness*

180 Level 1 comprises five major steps ([Box 2](#)). Step 1 verifies the presence and format of  
181 variables ensuring their inclusion and adherence to the designated format. Step 2 verifies  
182 missingness of values within the ConcePTION CDM and provides the extent of missing data  
183 overall and by calendar year. Step 3 verifies whether date variables comply with the required  
184 format and validates day, month, and year values. Steps 1 to 3 are performed simultaneously  
185 on all ConcePTION CDM tables, except for the METADATA table. The results are  
186 consolidated into a single HTML report.

<b>Box 2. Details on the quality checks involved in level 1: step 0 to 5</b>		
		<b>HTML report</b>
<b>Step 0: METADATA</b>	Checks the working directory and the consistent presence of all CDM tables and variables, irrespective of uppercase/lowercase and content	Yes
	Checks if all variables are written in lowercase.	
<b>Step 1: Checks the table formatting</b>	Checks if all rows of the .csv files in the working directory contain the correct number of fields	Yes
	Checks if all variables are present irrespective of their content	
	Checks if variables are written in lowercase	
	Checks for presence of mandatory variables as according to the CDM	
	Checks for presence of non-mandatory variables by comparing between the table of interest and the information recorded in the METADATA table	
	Checks presence of vocabularies for specific variables	
	Assess the format for all values and compare to a list of acceptable formats which have been filled out in the METADATA table	
<b>Step 2: Missing data analysis</b>	Tabulate missingness in all variables, overall and by calendar year (in the tables that contain a date variable)	
	Missing data will be further stratified by data provenance* (in the tables that containing this variable)	
	Missing data overall will be displayed using bar charts for each CDM table and reported as counts and percentages	
	Missing data stratified by data provenance or calendar year will be displayed using line charts for each CDM table and reported as counts and percentages	
	Missing data stratified by data provenance and calendar year will be displayed using faceted bar charts for each CDM table and reported as counts and percentages	
<b>Step 3: Date variables</b>	Checks if dates are in the correct format (8 characters)	
	Checks if date variables contain allowable values: Year: 1995(study specific)-present; Month: 01-12; Day: 01-31	
<b>Intermediate step</b>	High-level description of the CDM instance based on the information collected from CDM_SOURCES and INSTANCE tables	Yes
<b>Step 4: Convention and counts of categorical variables</b>	Checks if the table of interest contains any duplicate rows	Yes, per each CDM table
	Checks that all conventions for the table of interest have been adhered to	
	Constructs frequency tables of categorical variables, overall and by calendar year (when the table of interest contains a date variable).	
<b>Step 5: Distribution of continuous variables and dates</b>	For continuous variables mean, median, interquartile range, skewness and kurtosis will be reported.	Yes, per each CDM table
	Distribution of date variables will be reported as counts of dates overall and by calendar year.	
* data provenance- is defined by using the meaning variable in the CDM tables		

188 An intermediate step before Steps 4 to 5 generates a high-level characterization of the  
189 ConcePTION CDM instance by extracting information from the CDM\_SOURCE and  
190 INSTANCE tables, highlighting the key attributes of the data instance. A verification of  
191 uniqueness of records is also performed as duplicates check. Step 4 verifies compliance of  
192 the ConcePTION CDM tables with the required conventions. These conventions are specific  
193 to each CDM table and refer to the table structure and vocabulary rules. Moreover, frequency  
194 tables are constructed for categorical variables, providing a summary of values. Step 5  
195 produces the distribution of continuous variables and date variables. For continuous variables,  
196 descriptive statistics such as mean, median, interquartile ranges are calculated. Steps 4 and  
197 5 are performed separately for each CDM table. An HTML report is generated for each table,  
198 providing a separate overview of the results. All issues identified need to be fixed by updating  
199 the ETL design, or when not possible will be noted. In summary, Level 1 DQA aims to verify  
200 format coherence and completeness of each variable across the 15 CDM tables,  
201 encompassing a total of 356 individual indicators. The code is available at  
202 <https://github.com/UMC-Utrecht-RWE/INSIGHT-Level1>

203 *Level 1b: Data characterization at value level*

204 Level 1b relies on the successful conformance to the ConcePTION CDM specifications and  
205 consists of two steps ([Box 3](#)). In this level, various aspects related to the data's coherence  
206 and precision are evaluated through 26 checks ([Box 3](#)), including alignment of ATC codes at  
207 the same pharmacological level, or consistency of treatment duration expression (days or  
208 weeks) if available, among others. Level 1b does not generate an HTML report, but a set of  
209 csv files for each CDM table. The code is available at [https://github.com/UMC-Utrecht-](https://github.com/UMC-Utrecht-RWE/INSIGHT-Level1b)  
210 [RWE/INSIGHT-Level1b](https://github.com/UMC-Utrecht-RWE/INSIGHT-Level1b)

<b>Box 3. Details on the quality checks involved in level 1b</b>		
		<b>HTML report</b>
<b>Step 1</b>	Creates frequency tables for each variable within each CDM table by displaying counts of all values present.	No
<b>Step 2</b>	Creates contingency tables for categorical variables within each CDM table and shows the distribution of one variable across the categories of other variables.	No



211 *Level 2: Data Relational logic characterization*

212 Level 2 consists of 8 steps ([Box 4](#)) with a total of 57 indicators and verifies temporal plausibility  
 213 of date variables (Steps 2.1-2.3, 2.5-2.6) and consistency of encounter records across  
 214 different tables (Steps 2.4, 2.7-2.8). For example, step 2.1 examines if health encounters  
 215 occurred before an individual's date of birth, ensuring the chronological integrity of the data.  
 216 Similarly, Step 2.4 verifies that person IDs with a record in any of the CDM tables are also  
 217 listed in the PERSONS table. The R-script is available at [https://github.com/UMC-Utrecht-](https://github.com/UMC-Utrecht-RWE/INSIGHT-Level2)  
 218 [RWE/INSIGHT-Level2](https://github.com/UMC-Utrecht-RWE/INSIGHT-Level2)

<b>Box 4. Details on the quality checks involved in level 2: step 1 to 8</b>		
		<b>HTML report</b>
<b>Step 1*</b>	Checks dates of events before birth	Yes
<b>Step 2*</b>	Checks dates of events after death	Yes
<b>Step 3*</b>	Checks dates of events outside observation periods	Yes
<b>Step 4*</b>	Checks subjects observed in a CDM table of interest without a corresponding record in the PERSONS table	Yes
<b>Step 5</b>	Checks observations associated with a <i>visit_occurrence_id</i> which occur before the <i>visit_start_date</i> (optional, if the study cares of visits)	Yes
<b>Step 6</b>	Checks observations associated with a <i>visit_occurrence_id</i> which occur after the <i>visit_end_date</i> (optional, if the study uses variable visits)	Yes
<b>Step 7</b>	Checks observations associated with a <i>visit_occurrence_id</i> for which the associated <i>person_id</i> differs from that in the VISIT_OCCURRENCE table (optional)	Yes
<b>Step 8</b>	Checks for subjects indicated in PERSON_RELATIONSHIPS as the parent of a child whose birth date is less than 12 years prior to the recorded birth date of the linkage child (optional, if the study requires mother-child linkage)	Yes
*Mandatory steps by default in the pipeline. Other steps could become mandatory based on study requirements.		

219 *Level 3: Data Content Characterization*

220 Level 3 creates tables and graphics for study-variables ([Box 5](#)) in 9 steps. First the population,  
 221 its follow-up, and entry and exit patterns are described, allowing for verification of stability of  
 222 the dynamic population. Subsequently key study variables such as medicines, vaccines,  
 223 diagnoses, pregnancy, lifestyle factors and EUROCAT indicators are created in the study  
 224 population. All indicators result in graphics allowing to inspect temporal trends, including  
 225 person-time, prevalence of medicine prescriptions, and incidence of events for the study  
 226 population and subpopulations of interest. This allows for the identification of any peaks,  
 227 drops, or trends in the data and facilitates benchmarking between different instances, DAPs

228 and external data. There are three versions of the output provided, one with no masking  
 229 procedures applied, the two other versions allow for masking of cell counts. The first is  
 230 masking all counts lower than five, and the second does not show absolute counts but only  
 231 ranges. With over 249 indicators, Level 3 allows to deeply assess whether data is fit for  
 232 purpose. The code is available at <https://github.com/UMC-Utrecht-RWE/INSIGHT-Level3>

<b>Box 5. Details on the quality checks involved in level 3: step 1 to 9</b>		
		<b>HTML report</b>
<b>Step 0: Subpopulation availability*</b>	Checks presence or absences of subpopulation* for iterative performance of level checks	No
<b>Step 1: Study-source population</b>	Checks counts, rates, and distributions of study population characteristics	Yes
<b>Step 2: Dates</b>	Checks the distribution of start-end, follow-up, and birth dates across time	Yes
<b>Step 3: Visits and lifestyle</b>	Checks counts and rates about lifestyle factors and use of healthcare facilities	Yes
<b>Step 4: Medicines</b>	Checks counts and rates across time of medicine use	Yes
<b>Step 5: Vaccines</b>	Checks counts and rates across time of vaccine exposure	Yes
<b>Step 6: Diagnoses</b>	Checks counts and rates across time of events of interest	Yes
<b>Step 7: Pregnancy</b>	Checks counts and rates of pregnant population stratified by status of pregnancy	Yes
<b>Step 8: Population of interest</b>	Checks counts and rates of events, medicine use and vaccine exposure distribution in a specific sub-group (e.g., pregnant women, subjects with a migraine diagnosis etc.)	Yes
<b>Step 9: EUROCAT**</b>	Data Quality Indicators on congenital anomalies information in Europe	Yes

\* Subpopulations refer to subsets of the source population and can be defined by the data provenance (i.e., hospital admission and general practitioner' data etc.)

\*\* The number of quality checks and HTML reports will double per the number of subpopulations available on the data instance For DAPs with EUROCAT/EUROmediCAT data sources, steps 1 and 2 are mandatory.

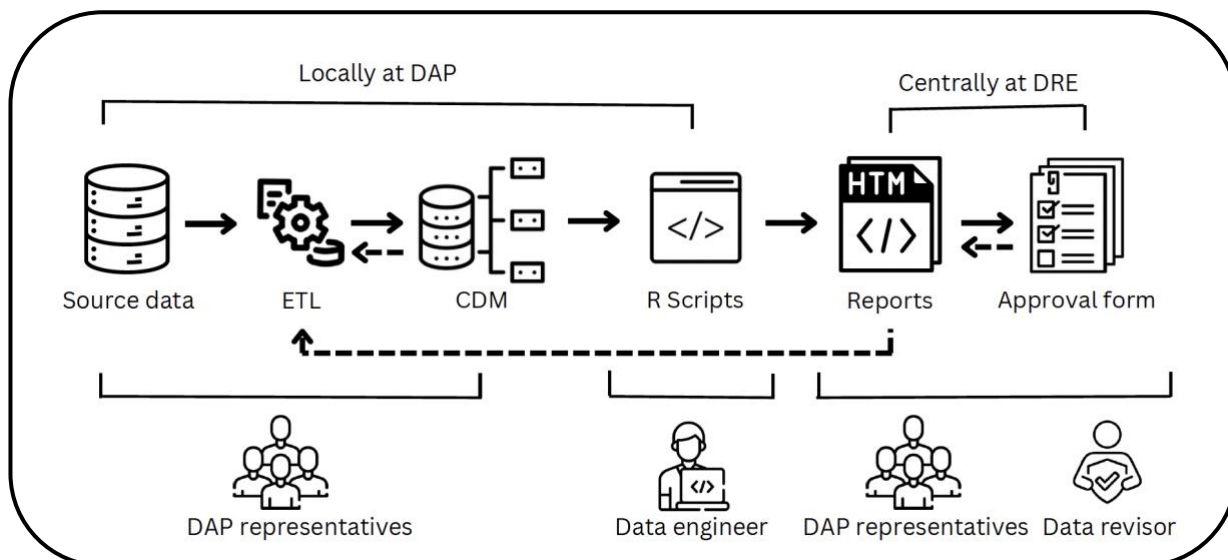
### 233 **HTML report**

234 HTML reports are generated for Level 1, 2 and 3 outputs. Within Level 1 several reports are  
235 created. The first report lists the presence of the METADATA table, a mandatory table, in the  
236 working directory. The report on Step 1 to 3 provides information on the format coherence and  
237 completeness. Reports Step 4 to 5 are optional and based on study specific requirements of  
238 CDM tables ([Box 1](#)). Reports 1-4 in Level 2 are mandatory, whereas reports 5-8 are optional,  
239 depending on study requirements ([Box 3](#)). Level 3 DQAs produce nine reports of which report  
240 1 and 2 are mandatory. The total number of reports depends on the study requirements and  
241 the number of subpopulations ([Box 5](#)). The output of level 1b is available as csv files and  
242 makes it possible to highlight any inconsistencies or create dictionaries for semantic  
243 harmonization of non-harmonized values.

### 244 **Responsible Parties**

245 The R-scripts are available on the UMCU-RWE private GitHub and maintained by UMC  
246 Utrecht. A public frozen version is available for this paper. DAPs are responsible for  
247 downloading and running the R scripts locally and addressing issues related to table formatting  
248 that results from Level 1 Step 1 to 3 and exceeding the *a-priori* 5% inconsistencies threshold  
249 defined in Level 2. If any issues are reported, DAPs must revise the ETL of the data instance  
250 and rerun these steps.

251 The outputs of the Level 1, 1b, 2 and 3 DQAs are uploaded to the UMCU Digital Research  
252 Environment (DRE) by the DAPs, for centralized review by the data quality revisor and the  
253 DAPs.<sup>14</sup> The data quality revisor collaborates with each DAP to address any issues flagged  
254 during the revision, as DAPs can provide background information on data collection  
255 processes, population characteristics, national healthcare system, availability of healthcare  
256 facilities, regulatory aspects, and expected disease incidence and prevalence. Observations  
257 are listed in the data quality approval form (Supplemental File 1). Data analytics engineers  
258 provide technical support related to issues that get reported. [Figure 2](#) shows the workflow.



**Figure 2.** Workflow of INSIGHT

259 **Approval form**

260 An approval form must be completed centrally by the DAP and the data quality revisor, with  
261 one form per instance and per level check. This form serves as a quality assurance document  
262 and systematically and transparently consolidates relevant information from the DQAs. The  
263 form is an Excel document with four tabs, one per level check listing the relevant aspects to  
264 be assessed per output (Supplemental File 1).

265 **Training**

266 As VAC4EU utilizes the ConcePTION CDM pipeline for its vaccine evaluation studies, they  
267 have developed a training course dedicated to Data Quality checks. Investigators can register  
268 and complete this course online. Upon successfully passing the final test, they will be awarded  
269 certification.

270 **ILUSTRATIVE EXAMPLE**

271 INSIGHT is available as a set of executable R scripts, currently supporting the use of  
272 ConcePTION CDM version 2.2.<sup>15–18</sup>

273 **Level 1**

274 In Level 1, one of the key objectives is to assess conformance of the ETL'ed data to the  
275 ConcePTION CDM specifications. For instance, if alternative values such as “1” for males and  
276 “2” for females are used instead of the designated values (M, F, U, O) for the variable  
277 *sex\_at\_instance\_creation* in the PERSONS table, it will be flagged as an error in Steps 1 to 3  
278 report.

279 Additionally, Level 1 identifies unallowable values in date variables. In [Figure 3](#), we show an  
280 example from the EVENTS table, the indicator *error\_year* is displayed in blue, highlighting  
281 records with dates before 1995 or in the future. Most DAPs have data banks starting after  
282 1995, that is why this cut off was chosen. The indicator *future\_dates* show records with dates  
283 in the future in comparison to the script running date.

284 DAPs are expected to resolve these issues by updating their ETL specifications to adhere to  
285 the specified rules and requirements, if possible.

286 **Level 1b**

287 Level 1b allows users to visualize various aspects of the data, such as the composition of  
288 codes in the PROCEDURES table (e.g., number, number with letters, symbols like “\$” or “?”)  
289 and provides insights into the original value definitions for study variables in the  
290 MEDICAL\_OBSERVATIONS or SURVEY\_OBSERVATIONS tables. Additionally, by  
291 assessing the content of the ConcePTION CDM tables the user can confirm that each table  
292 was fed appropriately.

293 **Level 2**

294 Step 2.3 highlights records where dates fall outside the observation period ([Figure 4](#)). The  
295 construction of this period is a crucial step. Start dates of observation can be obtained from  
296 various sources, usually inhabitant registries. However, accessibility to these sources may  
297 vary and alternative resources such as first visit to a general practitioner are used. In figure 4,  
298 the high percentages of dates outside the observation period indicate that the current  
299 observation period fails to capture all relevant events.

300 Moreover, Level 2 assesses the presence of subjects across the CDM tables. Ideally, all  
301 *person\_id* in the CDM tables should be in the PERSONS table. An error in step 4 indicates a  
302 need to re-evaluate the design of the PERSONS table.

303 **Level 3**

304 In Level 3, DQAs become more extensive, as their goal is to facilitate the evaluation of fitness-  
305 for-purpose. [Figure 5](#) illustrates a demographic tree indicating a similar distribution between  
306 females and males. However, the study focus of ConcePTION is on females and young males.  
307 It is possible that fathers have been included in the study population, which does not align with  
308 the study requirements but is not necessarily an error. Another example is gestational diabetes  
309 diagnostic codes in males. This could be explained by recording diagnoses under male  
310 children. However, this percentage should be low to reflect the reality of the condition,  
311 otherwise it may indicate an error.

312 In level 3, one of the checks is distribution analysis of start and end dates of follow-up, by year  
313 and month. In [Figure 6](#) all records start in January and July of 2004. This pattern indicates a  
314 potential ETL specification to filter out records outside the study period of interest. The peak  
315 in July suggests imputation due to missing data in date of birth, which has been used as start  
316 of observation.

317 Level 3 plays a crucial role in detecting and reflecting the consequences of technical errors  
318 during L1, L1b, and L2 and highlights any errors from the data generation processes. By  
319 identifying and addressing these issues, the goal is to improve the overall quality of the data  
320 source. INSIGHT is an iterative process of running and updating, until the DAPs and PI have  
321 confidence in the data source's quality.

**Figure 3.** Example of level 1 check

1. Table formatting
2. Dates check
3. Missing data analysis
4. Output folder structure

## 2. Dates check

Check if date variables have the correct format and contain only allowable values for year, month and day. Report if any of the databases contain dates in the future.

**Explanations:**

- a. If a date variable contains errors in its format, those dates will be removed from further analysis (allowable values check and counts of future dates).
- b. All errors are reported as counts of errors for each table in the directory.
- c. If a date variable does not contain any errors it will not be displayed in the table below.
- d. If error\_year is smaller than future\_dates that means that for those extra future dates the year of the date variable that contains the error is equal to the present year, but the day or month can be in the future.

**Allowable values:**

- a. year: 1995-present (in some cases such as: dates that show follow up, future dates are allowed but they will still be counted in error\_year)
- b. month: 01-12
- c. day: 01-31

The tables that will not undergo this analysis even if present in the directory due to absence of date variables are:

- a. PERSON\_RELATIONSHIPS
- b. PRODUCTS

Based on errors in future or past dates error\_year will be colored:

- a. **blue**: years before 1995 and in the future
- b. **red**: years before 1995 only
- c. **green**: years in the future only
- d. **black**: no errors

Show  entries

Search:

table_name	table_directory	variable_name	error_format	error_year
VISIT_OCCURRENCE	VISIT_OCCURRENCE_SPA201111.csv	visit_start_date	0	1
EVENTS	EVENTS_EXE.csv	start_date_record	0	4922
EVENTS	EVENTS_PS.csv	start_date_record	0	1
EVENTS	EVENTS_PS.csv	end_date_record	0	1
EVENTS	EVENTS_SALM.csv	start_date_record	0	45990
MEDICINES	MEDICINES_2005_SPF.csv	date_dispensing	0	49
PROCEDURES	PROCEDURES_SDO2003.csv	procedure_date	0	1
PROCEDURES	PROCEDURES_SDO2005.csv	procedure_date	0	2
PROCEDURES	PROCEDURES_SDO2007.csv	procedure_date	0	2
PROCEDURES	PROCEDURES_SDO2008.csv	procedure_date	0	8

Showing 1 to 10 of 39 entries

Previous  2 3 4 Next

**Figure 4.** Example of level 2 check

### 2.3 - Date values outside observation periods

The table below shows the number of date values taking place outside of the observation periods as specified in the OBSERVATION\_PERIODS table, by table, date variable, and meaning:

Show  entries Search:

TABLE	v.DATES	Meaning	Issues	Total	Percentage
EVENTS	start_date_record	(NOT ENTERED)	4578	42528	10.9
EVENTS	start_date_record	access_to_mental_health_service_comorbidity	580	6784	8.5
EVENTS	start_date_record	access_to_mental_health_service_primary	95365	826076	11.5
EVENTS	start_date_record	emergency_room_diagnosis	856234	8991024	9.9
EVENTS	start_date_record	emergency_room_presentation	856234	8991024	9.9
EVENTS	start_date_record	exemption	144673	1425621	10.1
EVENTS	start_date_record	hospitalisation_primary	1585712	6862820	23.1
EVENTS	start_date_record	hospitalisation_secondary	1071344	5318087	20.0
MEDICAL_OBSERVATIONS	mo_date	measure_during_hospitalisation	6054	43816	13.8
MEDICINES	date_dispensing	dispensing_in_community_pharmacy	13766879	200060691	6.9
MEDICINES	date_dispensing	dispensing_in_hospital_pharmacy_unspecified	2741080	38346942	7.4
MEDICINES	date_prescription	dispensing_in_community_pharmacy	0	0	0.0
MEDICINES	date_prescription	dispensing_in_hospital_pharmacy_unspecified	0	0	0.0
PROCEDURES	procedure_date	italian_outpatient	46972216	505311853	9.3
PROCEDURES	procedure_date	procedure_during_hospitalisation	2781140	14008393	19.9
SURVEY_ID	survey_date	birth_registry_mother	333694	1046685	31.8
SURVEY_ID	survey_date	death_registry	6753	37824	17.8
SURVEY_ID	survey_date	induced_termination_registry	98733	105625	93.0
SURVEY_ID	survey_date	spontaneous_abortion_registry	37807	64708	58.5
VACCINES	vx_admin_date	administration_of_vaccine_unspecified	324401	4088683	7.9
VACCINES	vx_record_date	administration_of_vaccine_unspecified	324401	4088683	7.9

Showing 1 to 21 of 21 entries Previous  Next



**Figure 5. Example of level 3 check**

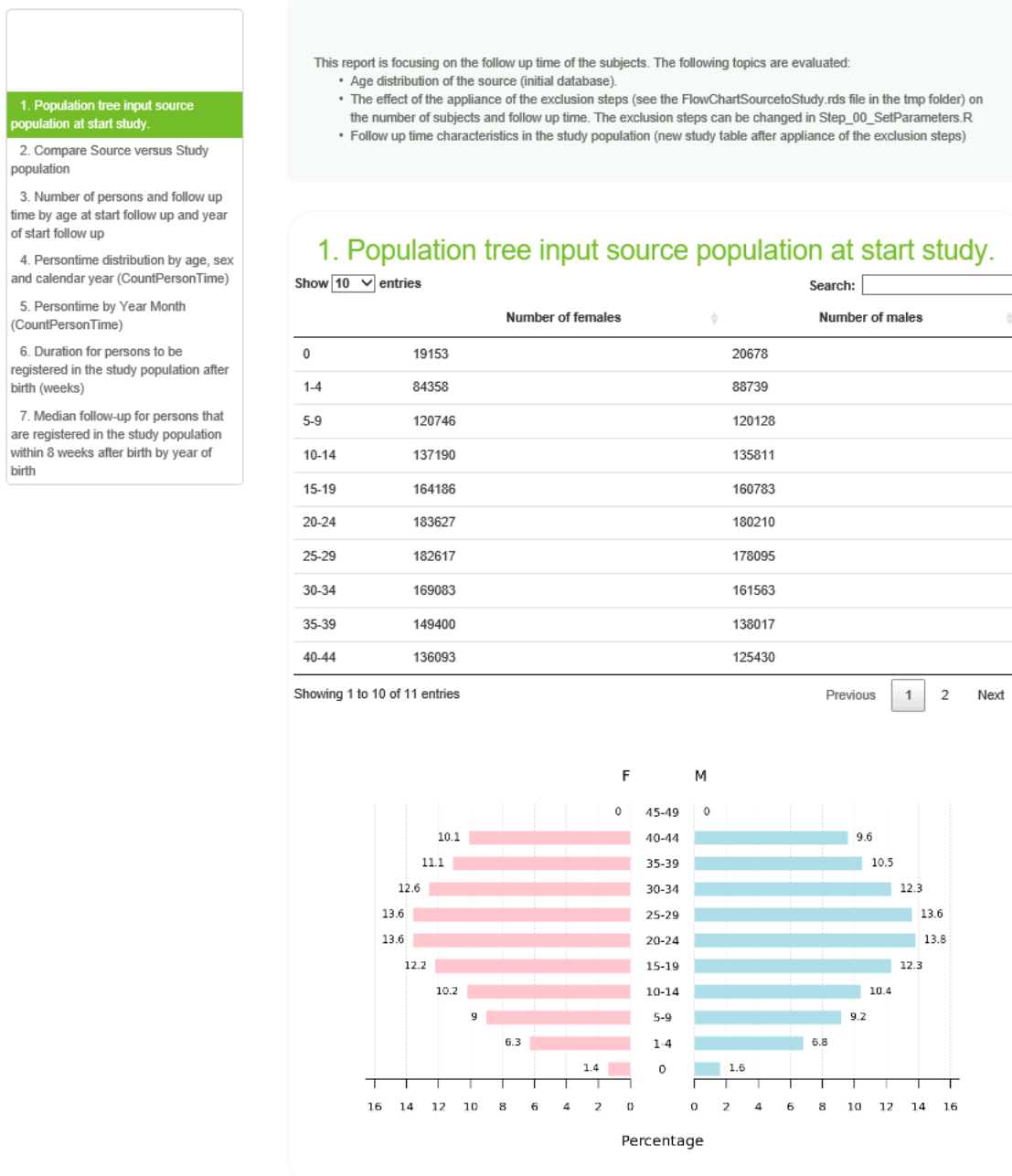
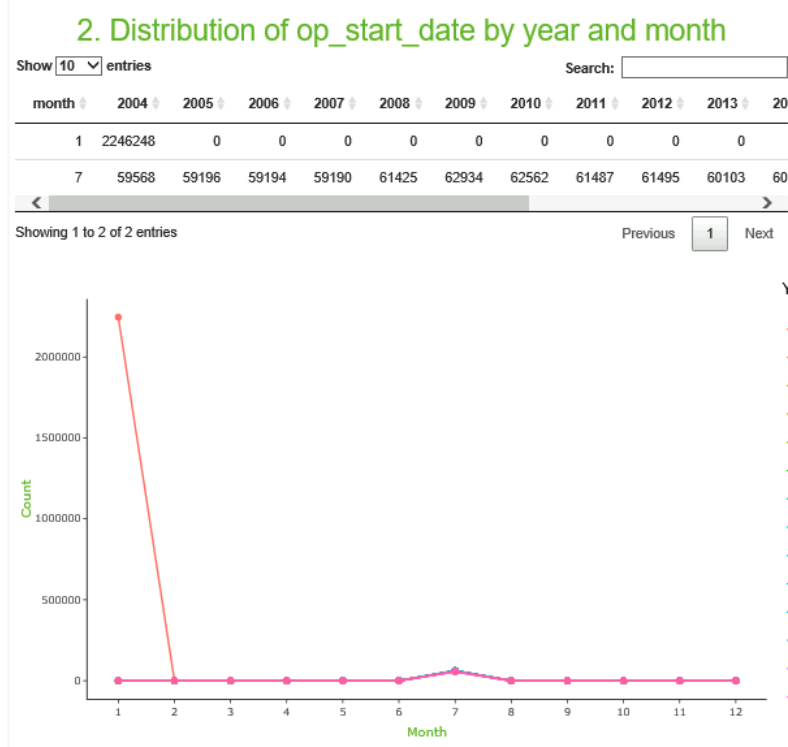


Figure 6. Example of level 3 check

- 1. Distribution of start\_follow\_up by year and month
- 2. Distribution of op\_start\_date by year and month
- 3. Distribution of end\_follow\_up by year and month
- 4. Distribution of op\_end\_date by year and month
- 5. Distribution of birth date by year and month
- 6. Distribution of birth date by day month



## 322 **DISCUSSION**

323 The use of RWD in research and decision-making has gained significant attention from  
324 regulatory agencies such as FDA, PMDA, and EMA. They have emphasized the importance  
325 of valid and trustworthy RWD for generation of RWE.<sup>1,2</sup> In this article, we present INSIGHT, a  
326 tool to assess fit-for-use (research independent) and fit-for-purpose (research specific) of  
327 RWD.

328 Studies including multiple data sources are becoming more popular especially in  
329 pharmacoepidemiologic studies.<sup>19</sup> With the abundance of healthcare data available, DAPs can  
330 leverage various data sources including routine healthcare data, surveillance systems,  
331 inhabitant registries, and more. However, the utility of these granular data sources relies on  
332 their quality. Data quality encompasses several features, including representativeness,  
333 precision, and accessibility, among other dimensions. To evaluate quality, EMA and Heads of  
334 Medicines Agency have proposed the DQF for EU medicines regulation which includes five  
335 quality dimensions: reliability, extensiveness, coherence, timeliness, and relevance.<sup>20</sup> We  
336 have mapped INSIGHT's data quality dimensions to Kahn's framework (left), and the EMA  
337 DQF (right) (Supplemental File 2). INSIGHT is built on the ConcePTION CDM, which is utilized  
338 by multiple networks (ConcePTION project, VAC4EU, EU & PV network, SIGMA consortium)  
339 and in many regulatory required multi-data source studies commissioned or requested by  
340 EMA.

341 CDMs play a vital role in facilitating data quality assessment when using secondary-use data.<sup>21</sup>  
342 While DQAs aid assessing fitness-for-purpose, it is essential to recognize that data quality is  
343 influenced by numerous factors, mostly the health care system, the purpose of data recording  
344 and the type of data recorder. INSIGHT DQAs provide a generic and transparent overview of  
345 data quality indicators, aiming to identify inconsistencies in the data, which can be categorized  
346 into three main types: extraction or convention errors, script errors, and patterns for  
347 acknowledgment and documentation. The outputs support the study team to understand the  
348 data that is being utilized and the semantic harmonization. The purpose of DQAs varies across  
349 these contexts, ranging from rectifying extraction errors to fine-tuning scripts to suit project  
350 requirements. Additionally, documenting data patterns ensures accurate interpretation and  
351 understanding of data nuances and limitations.

### 352 **Previous data quality assessment tools**

353 Various research networks have developed their own DQA processes with varying levels and  
354 indicators. For example, the US FDA Sentinel uses a four-level approach focusing on  
355 completeness and validity, accuracy and integrity, consistency of trends, and plausibility.<sup>9,22</sup>  
356 The EUROCAT network, has a comprehensive DQA process, that includes indicators such as

357 case ascertainment, accuracy of diagnoses, data completeness, timeliness, and availability of  
358 denominator data.<sup>23,24</sup> Researchers using the OMOP CDM have access to tools like the White  
359 Rabbit and the Data Quality Dashboard (DQD) for assessing data quality at source data and  
360 CDM-standardized database.<sup>8,25–27</sup>

361 A comparison of DQAs across six research networks revealed variations in the number of  
362 quality checks, ranging from 875 up to 3234.<sup>28</sup> The DQAs processes also varied, including  
363 differences in centralization, distributed coordination of DQAs, programming languages, and  
364 staff involvement. These differences can be attributed to specific network requirements,  
365 analytical focus, as well as the maturity of their DQA models.

### 366 **Strengths and limitations**

367 Our open-source DQA methodology was built during the ConcePTION project and has been  
368 made generic and is adopted by other networks using the ConcePTION CDM. Our approval  
369 system is based on a flexible approval form and relies on study requirements, over  
370 predetermined conditions. However, there are limitations to be acknowledged. First, script  
371 running time depends on hardware, data size and complexity. Components like the central  
372 processing unit, random access memory, and computer data storage help determine process  
373 efficiency. Second, the pipeline requires iterative reruns after data issues are fixed until  
374 desired quality is achieved or at least, explained. Steps must be executed sequentially to avoid  
375 downstream issues. Last, fit-for-purpose evaluation is time and resource-intensive. While  
376 summary reports and dashboards aid assessment, automated detection of trends and  
377 verification against external benchmarks need to be facilitated further. Automating outlier and  
378 pattern detection is an area for future development given the increase complexity and limited  
379 skilled human resources. Overall, the DQA workflow is subject to continuous improvements.

### 380 **Implications for researchers and future research**

381 The INSIGHT R scripts are openly available through GitHub, along with a training course for  
382 the Vaccine Monitoring Collaboration for Europe (VAC4EU) and educational videos to  
383 encourage other stakeholders to adopt this DQA pipeline and adhere to principles of open  
384 science. This tool has been already successfully implemented in EMA tendered risk  
385 minimization studies (EUPAS31095, EUPAS21001), COVID vaccines effectiveness study  
386 (EUPAS40404, EUPAS42504), Post-Authorization Safety Studies (EUPAS43593,  
387 EUPAS44424, EUPAS41725, EUPAS45461), and CONSIGN (EUPAS39226, EUPAS39438,  
388 EUPAS40317).<sup>29</sup> As previously stated, INSIGHT is compliant with the proposed DQF for EU  
389 medicines regulation launched in October 2022.

390 Interoperability between CDMs and their specific DQAs processes remains a challenge, as  
391 each CDM must be used in conjunction with its customized tools. This implies that applying

392 external DQAs tools to an already CDM-standardized data source would require an additional  
393 ETL process tailored to that CDM. To address this, we propose to establish standards and  
394 guidelines for achieving interoperability across different CDMs. This would facilitate the  
395 exchange of DQAs tools and methodologies between research networks and regulatory  
396 bodies, streamlining the quality assessment process.

### 397 **CONCLUSION**

398 In this article, we introduced INSIGHT, a comprehensive pipeline for conducting DQAs of  
399 ConcePTION CDM-standardized datasets. The tool aims to facilitate the evaluation of the  
400 fitness-for-use and fitness-for-purpose of RWD. By implementing this tool, researchers can  
401 assess the adequacy and reliability of the data, ensuring that the evidence generated is robust  
402 and reliable for studies on medicines and vaccines. INSIGHT is a valuable tool for evaluating  
403 the quality of data sources and enhancing the credibility of RWD in clinical and regulatory  
404 decision-making.

## REFERENCES

1. U.S. Food and Drug Administration. Framework for FDA's Real-World Evidence Program. Published 2023. Accessed February 1, 2023. <https://www.fda.gov/media/120060/download>
2. Heads of Medicines Agency. European Medicines Agency HMA-EMA Joint Big Data Taskforce Summary Report. Accessed February 1, 2023. [https://www.ema.europa.eu/en/documents/minutes/hma/ema-joint-task-force-big-data-summary-report\\_en.pdf](https://www.ema.europa.eu/en/documents/minutes/hma/ema-joint-task-force-big-data-summary-report_en.pdf)
3. Trifirò G, Coloma PM, Rijnbeek PR, et al. Combining multiple healthcare databases for postmarketing drug and vaccine safety surveillance: why and how? *J Intern Med*. 2014;275(6):551-561. doi:10.1111/joim.12159
4. Togo K, Yonemoto N. Real world data and data science in medical research: present and future. *Japanese Journal of Statistics and Data Science*. 2022;5(2):769-781. doi:10.1007/s42081-022-00156-0
5. Illamola SM, Bucci-Rechtweg C, Costantine MM, Tsilou E, Sherwin CM, Zajicek A. Inclusion of pregnant and breastfeeding women in research - efforts and initiatives. *Br J Clin Pharmacol*. 2018;84(2):215-222. doi:10.1111/bcp.13438
6. Bromley R, Weston J, Adab N, et al. Treatment for epilepsy in pregnancy: neurodevelopmental outcomes in the child. *Cochrane Database of Systematic Reviews*. 2014;2020(6). doi:10.1002/14651858.CD010236.pub2
7. Qualls LG, Phillips TA, Hammill BG, et al. Evaluating Foundational Data Quality in the National Patient-Centered Clinical Research Network (PCORnet®). *eGEMs (Generating Evidence & Methods to improve patient outcomes)*. 2018;6(1):3. doi:10.5334/egems.199
8. Blacketer C, Defalco FJ, Ryan PB, Rijnbeek PR. Increasing trust in real-world evidence through evaluation of observational data quality. *Journal of the American Medical Informatics Association*. 2021;28(10):2251-2257. doi:10.1093/jamia/ocab132
9. Ball R, Robb M, Anderson S, Dal Pan G. The FDA's sentinel initiative-A comprehensive approach to medical product surveillance. *Clin Pharmacol Ther*. 2016;99(3):265-268. doi:10.1002/cpt.320
10. Thurin NH, Pajouheshnia R, Roberto G, et al. From Inception to ConcePTION: Genesis of a Network to Support Better Monitoring and Communication of Medication Safety During Pregnancy and Breastfeeding. *Clin Pharmacol Ther*. 2022;111(1):321-331. doi:10.1002/cpt.2476
11. Kahn MG, Callahan TJ, Barnard J, et al. A Harmonized Data Quality Assessment Terminology and Framework for the Secondary Use of Electronic Health Record Data. *eGEMs (Generating Evidence & Methods to improve patient outcomes)*. 2016;4(1):18. doi:10.13063/2327-9214.1244
12. Bian J, Lyu T, Loiacono A, et al. Assessing the practice of data quality evaluation in a national clinical data research network through a systematic scoping review in the era of real-world data. *Journal of the American Medical Informatics Association*. 2020;27(12):1999-2010. doi:10.1093/jamia/ocaa245
13. Sturkenboom M, Braeye T, van der Aa L, et al. ADVANCE database characterisation and fit for purpose assessment for multi-country studies on the coverage, benefits and risks of pertussis vaccinations. *Vaccine*. 2020;38:B8-B21. doi:10.1016/j.vaccine.2020.01.100
14. anDREa Consortium. anDREa. Accessed February 1, 2023. [www.andrea-cloud.eu](http://www.andrea-cloud.eu)

15. Hoxhaj V. UMC-Utrecht-RWE/INSIGHT-Level1: V1. Zenodo. 2023. doi: 10.5281/zenodo.10035167.
16. Cid Royo A. and Elbers R. UMC-Utrecht-RWE/INSIGHT-Level1b: V1. Zenodo. 2023. doi: 10.5281/zenodo.10048081.
17. Hoxhaj V, and van den Bor R. UMC-Utrecht-RWE/INSIGHT-Level2: V1. Zenodo. 2023 doi: 10.5281/zenodo.10035169.
18. Hoxhaj V, Elbers R, and Alsina E. UMC-Utrecht-RWE/INSIGHT-Level3: V1. Zenodo. 2023. doi: 10.5281/zenodo.10035171.
19. Pacurariu A, Plueschke K, McGettigan P, et al. Electronic healthcare databases in Europe: descriptive analysis of characteristics and potential for use in medicines regulation. *BMJ Open*. 2018;8(9):e023090. doi:10.1136/bmjopen-2018-023090
20. EMA and Heads of Medicines Agency. Data Quality Framework for EU medicines regulation. Accessed February 1, 2023. <https://www.ema.europa.eu/en/news/high-quality-data-empower-data-driven-medicines-regulation-european-union>
21. Trifirò G, Coloma PM, Rijnbeek PR, et al. Combining multiple healthcare databases for postmarketing drug and vaccine safety surveillance: why and how? *J Intern Med*. 2014;275(6):551-561. doi:10.1111/joim.12159
22. Behrman RE, Benner JS, Brown JS, McClellan M, Woodcock J, Platt R. Developing the Sentinel System — A National Resource for Evidence Development. *New England Journal of Medicine*. 2011;364(6):498-499. doi:10.1056/NEJMp1014427
23. European Commission. Data Quality Indicators in EUROCAT. Accessed February 1, 2023. <https://eu-rd-platform.jrc.ec.europa.eu/sites/default/files/DQI-List-of-Data-Quality-Indicators-since-2012.pdf>
24. Loane M, Dolk H, Kelly A, Teljeur C, Greenlees R, Denssem J. Paper 4: EUROCAT statistical monitoring: Identification and investigation of ten year trends of congenital anomalies in Europe. *Birth Defects Res A Clin Mol Teratol*. 2011;91(S1):S31-S43. doi:10.1002/bdra.20778
25. OHDSI. Data Quality Dashboard (DQD). Accessed February 1, 2023. <https://ohdsi.github.io/DataQualityDashboard/>
26. OHDSI. CDM OHDSI Common Data Model. Accessed February 1, 2023. <https://github.com/OHDSI/CommonDataModel/releases>
27. OHDSI. WhiteRabbit. Accessed February 1, 2023. <https://github.com/OHDSI/WhiteRabbit>
28. Callahan TJ, Bauck AE, Bertoch D, et al. A Comparison Of Data Quality Assessment Checks In Six Data Sharing Networks. *eGEMs (Generating Evidence & Methods to improve patient outcomes)*. 2017;5(1). doi:10.13063/2327-9214.1287
29. Maisonneuve E, de Bruin O, Hurley E, et al. CONSIG-International Meta-Analysis Protocol & Statistical Analysis Plan (1.1). Zenodo. doi:10.5281/zenodo.7112746

## **FUNDING**

The ConcePTION project has received funding from the Innovative Medicines Initiative 2 Joint Undertaking under grant agreement No 821520. This Joint Undertaking receives support from the European Union's Horizon 2020 research and innovation programme and EFPIA.

## **AUTHORS CONTRIBUTION**

VH, CD, and MCJM started with the initial design; VH, RE, EA, and RvdB programmed the R scripts; VH and CLAN wrote the initial draft of the manuscript. VH, CLAN, JRA, RE, EA and MCJM, provided critical feedback and edited the manuscript. All authors have seen and approved the manuscript. The research leading to these results was conducted as part of the ConcePTION consortium. This paper only reflects the personal views of the stated authors.

## **ACKNOWLEDGMENTS**

We would like to thank Rutger van den Bor for supporting the programming of INSIGHT Level 2, Claudia Bartolini for testing the scripts, and Rosa Gini for her expertise and collaboration throughout this project.

## **SUPPLEMENTARY MATERIAL**

**S1.** Approval form

**S2.** Theoretical mapping of INSIGHT to other DQFs

## **CONFLICT OF INTEREST STATEMENT**

All authors have completed the ICMJE uniform disclosure form at [www.icmje.org/coi\\_disclosure.pdf](http://www.icmje.org/coi_disclosure.pdf) and declare: all authors received financial support through the ConcePTION project for the submitted work; no financial relationships with any organizations that might have an interest in the submitted work in the previous three years; no other relationships or activities that could appear to have influenced the submitted work.

## **INSIGHT PIPELINE AVAILABILITY**

The quality check pipeline for ConcePTION CDM v2.2 instances is available as an open-source R scripts in the following link <https://github.com/UMC-Utrecht-RWE>. Protocol and Statistical Analysis Plan (SAP) can also be found there. Furthermore, a training course for DAPs and researchers is available upon request.

## **STUDY REGISTRATION**

This research was registered in EU PAS registration with number EU50142.